**EKONOMICKÁ UNIVERZITA V BRATISLAVE**

**Fakulta hospodárskej informatiky**

Logo

Description automatically generated

**Projekt z predmetu Big Data 02**

Bc. Richard Mišek

2023   Bc. Pavol Melko

Obsah

[Úvod 3](#_Toc132238362)

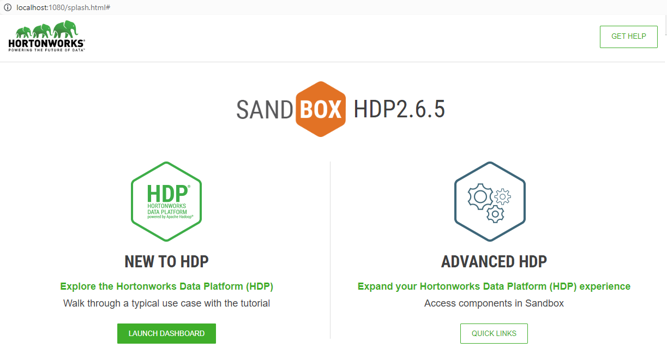
[Hortonworks 4](#_Toc132238363)

# Úvod

Úlohou tohto zadania je oboznámiť sa s Ambari, načítať doň dva CSV súbory a následne vytvoriť SQL dopyt na vytvorenú databázu. Obe tabuľky na seba musia nadväzovať. V našom prípade sme pracovali s tabuľkami, ktorých dáta boli z oblasti filmov a hodnotenia filmov. Hodnotenie spočívalo v spokojnosti divákov, kde súbor ratings obsahuje id hodnotiaceho, id filmu a príslušné priradené hodnotenie. Súbor anime obsahuje súpis všetkých filmov, ku ktorým diváci mohli priradiť hodnotenia.

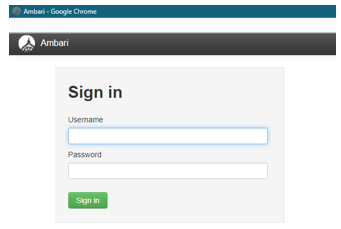
# Hortonworks

Nainštalovali sme rozsiahly súbor na virtual box a po úspešnej inštalácii sme dosiahli nasledovný výstup. Signalizuje nám úspešnú inštaláciu a následné spustenie Hortonworks Data Platform 2.6.5.



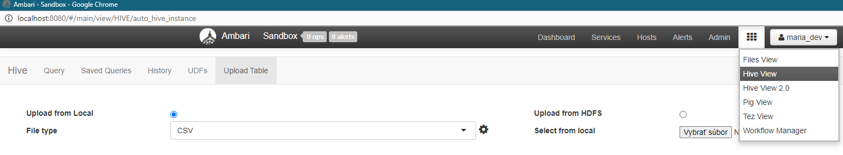
Pre vstúpenie do Ambari Dashboard sme stlačili zelené tlačidlo LAUNCH DASHBOARDS

Následne bolo treba vstúpiť pomocou prihlasovacích údajov do Ambari, kde username aj password sú maria\_dev

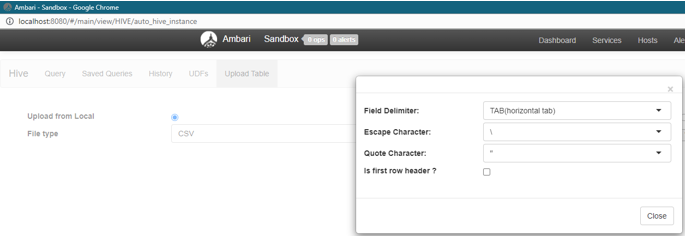


Pri práci s týmto zadaním bolo potrebné spustiť služby HDFS, YARN, MAPREDUCE, HIVE. V prípade ak spustené nie sú spustíme ich manuálne.

V pravej vrchnej lište po kliknutí na menu “štvorčeky“ a následne na Hive View. Po stlačení si zvolíme záložku Upload Table.



Následne sme cez ozubené koliesko nastavili požadované oddeľovače, aby sme z neskôr poskytnutých CSV súborov, vedeli preniesť dáta a pracovať s nimi v tabuľkách.

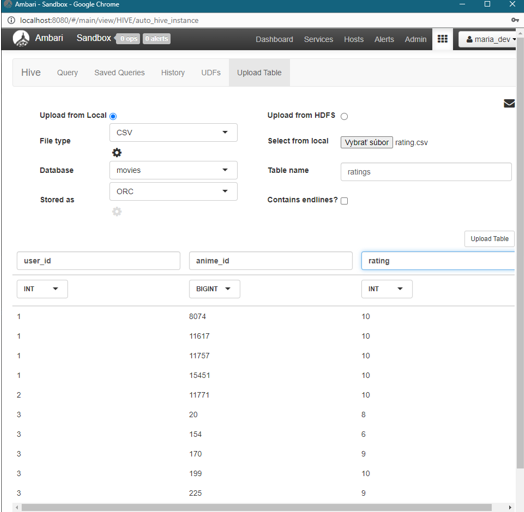


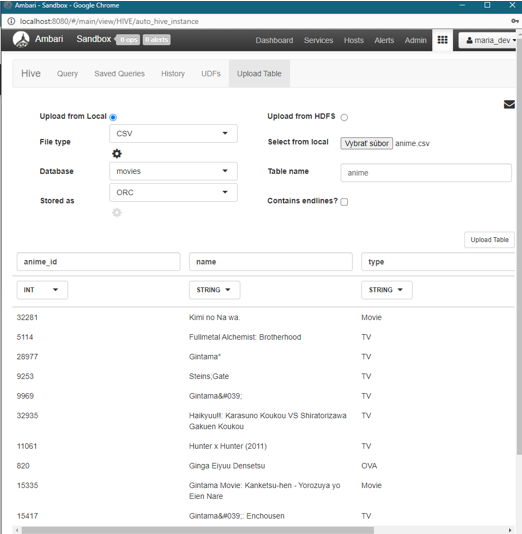
Najskôr sme si vytvorili novú databázu nazvanú movies

Obrázok, na ktorom je text

Automaticky generovaný popis

V tejto databáze sme následne vytvorili a vložili dve tabuľky : ratings a anime. Pri vytváraní sme dbali na dodržanie mena databázy, do ktorej sme vkladali súbory a správnosť typov súboru. Pri vytváraní sme museli zadať názvy atribútov a prvý riadok z CSV odstrániť, nakoľko tento riadok obsahoval názvy, ktoré sa premietali ako prvý riadok hodnôt. Môžeme si všimnúť, že pri anime\_id sme zvolili typ BIGINT. Voľba tohto dátového typu bola vykonaná vzhľadom na veľký obnos anime filmov v databáze. V prípade ak by sa filmy časom dopĺňali, nevznikali by problémy s kapacitou dátového typu.





Vidíme, že dáta sa nám preniesli z CSV súborov do databázy v poriadku.

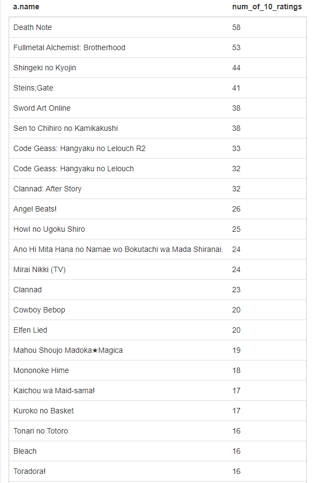
Súbory, s ktorými sme pracovali sú dostupné na : <https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database?select=anime.csv>

Prvý CSV súbor ratings, po úprave obsahuje 22 000 riadkov a súbor anime obsahuje 12 000 riadkov.

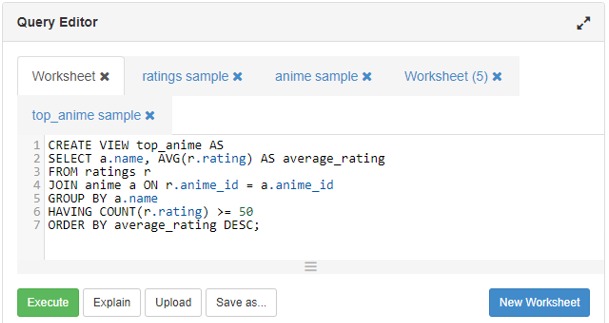
Po načítaní údajov a vytvorení tabuliek, sme overili funkčnosť dvoma SQL dopytmi na vytvorenú databázu.

Obrázok, na ktorom je text

Automaticky generovaný popis



Po vykonaní druhej Query, sme v databáze vytvorili novú tabuľku top\_anime, v ktorej sme uložili výstup. Týmto prístupom vieme dlhodobo uchovávať výstupy v databáze a následne s údajmi pracovať aj v budúcnosti.





Pri spracovávaní údajov nám nenastali žiadne väčšie komplikácie. SQL dopyty dopadli úspešne a výstupy boli vypísané. Jednotlivé výstupy sme overili cez MS Excel.